TRIZ 分析矩阵智能构建方法研究*

邢思思 [1,2]　　钱力 [1,2]

[1]（中国科学院文献情报中心 北京　100190）

[2]（中国科学院大学经济与管理学院图书情报与档案管理系 北京　100190）

摘要：[目的]基于深度学习和领域知识图谱技术，研究 TRIZ 分析矩阵自动构建方法，为技术创新提供知识支持。**[方法]**首先面向领域需求，将通用的“技术要素”延伸细化为实体和关系类别，完成领域知识图谱模式层设计；然后基于 BERT 预训练语言模型，研究设计从专利文献中自动识别知识实体及关系的智能化工具，实现实体关系自动抽取；最后利用图数据库完成领域知识图谱的构建，基于知识查询读取需要的知识实体和关系，实现 TRIZ 分析矩阵的自动构建。**[结果]**经实证验证，本文构建的面向薄膜磁头技术领域专利的 BERT-MH+CRF 实体识别模型 F1 分数为 84.93%，BERT-MH+softmax 关系抽取模型 F1 分数为 63.7%。**[局限]**缺乏知识推理技术的应用，无法揭示实体间潜在的知识关联，构建的 TRIZ 分析矩阵的质量仍存在不足。**[结论]**所提出的方法可以实现 TRIZ 分析矩阵自动构建的目标，可以为技术创新提供有效的知识支持。

关键词：TRIZ 分析矩阵；专利文献；预训练技术；实体识别；关系抽取

分类号：G250

Research on Intelligent Construction Method of the TRIZ Analysis Matrix

Xing Sisi[1,2], Qian Li[1,2]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [Objective] Based on deep learning and domain knowledge graph technology, study the automatic construction method of TRIZ analysis matrix to provide knowledge support for technological innovation. **[Methods]** First, facing the domain requirements, the general "technical elements" are extended and refined into entity and relationship categories, and the pattern layer design of domain knowledge graph is completed;Then, based on the BERT pre-trained language model, research and design an intelligent tool to automatically identify knowledge entities and relationships from patent documents, and realize the automatic extraction of entity relationships;Finally, the graph database is used to complete the construction of the domain knowledge graph, and the knowledge entities and relationships are required to read based on the knowledge query, so as to realize the automatic construction of the TRIZ analysis matrix. **[Results]** After empirical verification, the F1 score of the BERT-MH+CRF entity recognition model for patents in the field of thin-film magnetic head technology constructed in this paper is 84.93%, and the F1 score of the BERT-MH+softmax relation

* 基金项目：本文系科技部创新方法工作专项项目“基于群智理论的创新方法新系统研究与应用示范”（项目编号：2019IM020100）的研究成果之一。

extraction model is 63.7%. [Limitations] The lack of application of knowledge reasoning technology cannot reveal the potential knowledge associations between entities, and the quality of the constructed TRIZ analysis matrix is still insufficient. [Conclusions] The proposed method can achieve the goal of automatic construction of TRIZ analysis matrix, which can provide effective knowledge support for technological innovation.
Keywords: TRIZ Analysis Matrix, Patent Literature, Pre-training Techniques, Entity Recognition, Relation Extraction

1 引言

中国正从“制造大国”向“智造强国”战略转型，以科技创新为核心的创新驱动发展战略已上升为国家战略，国家和企业对创新的需求不断提高，主要体现在产品快速迭代的需求、技术交叉加剧的需求和创新知识集中汇聚的需求。在当前的时代背景下，创新已经不仅仅是依靠个人灵感而产生的想法，而更需要科学的方法和依据给予突破[1]。TRIZ 创新方法通过对专利大数据的挖掘分析，形成了一套指导人们进行发明创新的系统化的方法学体系，可以准确分析和发现核心问题，提高创新的效率与质量，是行之有效的创新方法之一。

系统相互作用分析矩阵是 TRIZ 创新方法体系中实现功能分析的核心技术。系统相互作用分析矩阵基于人工方法从专利文献中提取组件和功能两类知识实体及对应关系，再建立矩阵开展系统功能分析，指导技术创新。然而，面对全球科技快速进步迭代和专利申请量的爆炸式增长，传统的系统相互作用分析矩阵也面临知识要素不足、人工构建速度慢、人力成本高等问题，导致利用 TRIZ 创新方法促进技术创新的效应偏低，难以满足国家和企业日益增长的创新需求。

面对上述背景及发展需求，本研究在 TRIZ 创新方法的指导下，面向特定领域拓展系统相互作用分析矩阵涵盖的知识要素，同时将本文提出的面向特定领域且包含更丰富知识实体和关系的矩阵定义为 TRIZ 分析矩阵，并提出利用大数据和人工智能技术开展 TRIZ 分析矩阵智能构建方法研究。具体的，利用 BERT 预训练语言模型技术，从海量专利数据中自动识别知识实体和关系，完成领域知识图谱构建；面向具体需求，基于知识查询和知识推理，从领域知识图谱中自动读取所需要的知识实体和关系，实现 TRIZ 分析矩阵的自动构建。一方面解决 TRIZ 创新方法的知识供应问题，另一方面为 TRIZ 创新方法的持续发展和完善提供支持。

2 研究现状

本文以 TRIZ 分析矩阵智能构建为总目标，重点研究从专利文献中自动抽取知识实体和关系的技术解决方案。因此，本文主要面向 TRIZ 分析矩阵的知识抽取相关研究开展调研，系统调研了通用的科学知识抽取方法和面向专利的知识抽取方法。

2.1 科学知识抽取研究

面向科学文献的知识抽取方法经历了基于词典和规则[2-4]、基于统计机器学习[5-6]、基于深度学习[7-10]、基于预训练语言模型的方法以及综合性抽取的长足发展。2018 年，Google 的研究人员 Devlin 等[11]提出 BERT 模型，BERT 模型采用

双向的 Transformer Encoder 结构，面向大规模公开语料进行预训练，得到了表征能力更强的预训练字向量，极大地提高了模型的性能。文献[12]应用 BioBERT 从 2900 万篇 PubMed 摘要中抽取生物实体，取得了最优性能。Zhang 等[13]在中文临床语料库上对 BERT 进行了预训练，并将得到的词嵌入作为 BiLSTM-CRF 模型的输入特征，解决乳腺癌命名实体识别问题。文献[14]提出了一个既利用预训练的 BERT 语言模型又结合目标实体信息解决关系分类任务的模型，与最新方法相比取得了显著改进。唐晓波等[15]利用 BERT 预训练语言模型，搭建 BERT-BiGRU-CRF 标注序列模型，面向金融文本语料联合抽取实体关系。综上所述，“预训练+微调”技术通过加入通用有效的语言知识编码，极大地提升了实体关系抽取的性能表现。

2.2 面向专利的知识抽取研究

自 Tsourikov 等人的开创性工作以来[16]，专利知识抽取已经提出了多种方法，包括 SAO 方法、基于本体的方法、统计机器学习方法等。胡正银等[17]基于 SAO 结构语义分析与 LDA 聚类方法，面向专利文献识别通用的 TRIZ 技术信息（技术问题、解决方案、技术功能、技术效果）；H.B.Kim 等[18]基于 SAO 方法抽取专利中的技术问题和技术方案实体；李晓曼等[19]基于 SAO 方法面向纳米肥料领域专利文献完成材料、产品、方法、功效和用途 5 种实体的识别。面向专利文献抽取知识实体及关系的机器学习模型主要包括最大熵模型、SVM 模型和 CRF 模型。李卫超等[20]提出了一种基于词法分析、语法分析和最大熵分类模型的专利功能信息抽取方法；Nanba 等[21]基于 SVM 方法识别学术文献和专利中的技术和效果两类信息；赖英旭等[22]结合 TRIZ 理论设计了水稻育种方法本体结构，应用 SVM 模型识别专利中的育种方法，应用 CRF 模型识别水稻品种。

从上述调研结果中可以看出，面向 TRIZ 分析矩阵的知识抽取技术，主要是面向专利的实体识别和关系抽取技术还远没有成熟，仍存在缺乏标注数据集、抽取的专利知识类型不足、自动化程度低、性能有待提高等问题。因此，本文拟利用预训练语言模型技术改进面向专利的知识实体识别和关系抽取。

3 TRIZ 分析矩阵智能构建方法设计

本研究提出的 TRIZ 分析矩阵智能构建方法框架如图 1 所示，主要由三个模块构成：面向 TRIZ 分析矩阵的领域知识图谱模式层设计、面向专利的知识实体和知识关系抽取方法、基于领域知识图谱的 TRIZ 分析矩阵自动构建。

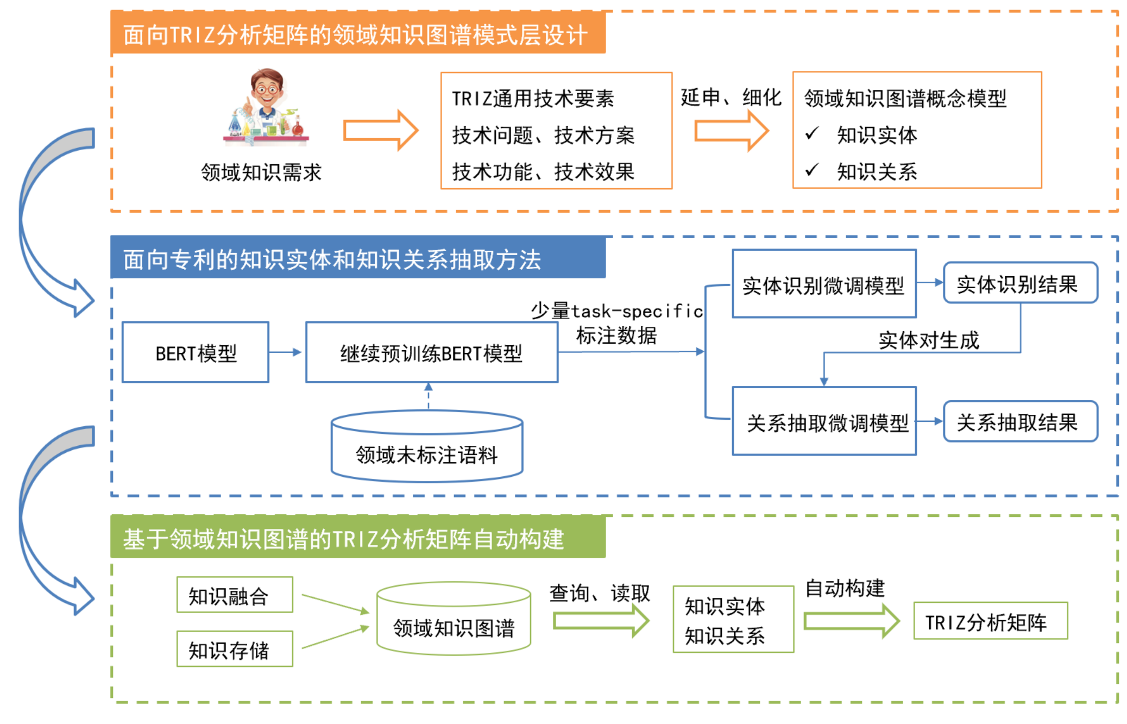


图 1　TRIZ 分析矩阵智能构建的方法框架

方法流程如下：（1）在 TRIZ 创新理论的指导下，面向特定领域的知识需求，将语义 TRIZ 中通用的“技术要素”（技术问题、技术方案、技术功能、技术效果）延伸细化为特定领域的知识实体和关系类别，完成领域知识图谱模式层设计。（2）基于 BERT 预训练语言模型，研究设计从专利文献中自动识别 TRIZ 分析矩阵构建所需要的知识实体及关系的智能化工具，实现专利实体和关系的自动抽取。（3）对知识实体和关系进行简单融合，利用图数据库完成领域知识图谱的构建；面向具体需求，基于知识查询和知识推理，从领域知识图谱中自动查询和读取所需要的知识实体和关系，实现 TRIZ 分析矩阵的自动构建。

上述 TRIZ 分析矩阵智能构建方法框架提供了全领域通用的方法流程和模型架构，明确了领域知识图谱模式层设计流程，搭建了通用的继续预训练 BERT 模型、实体识别微调模型和关系抽取微调模型架构，提供了标准化的领域知识图谱构建方法。在面向特定领域开展应用研究时，只需提供领域需求和领域专利语料，即可在上述方法框架的指导下开展 TRIZ 分析矩阵构建。

3.1　面向 TRIZ 分析矩阵的领域知识图谱模式层设计

语义 TRIZ 利用语义技术自动或半自动建模专利中隐含的技术信息，可以有效表示“技术问题、技术方案、技术功能、技术效果”等专利中特有的技术知识[23]。本文参考语义 TRIZ 模型，将 TRIZ 分析矩阵知识模型中的知识要素明确为技术问题、技术方案、技术功能、技术效果 4 大功能语义类型，开展领域知识图谱模式层设计，具体流程如图 2 所示。首先，面向特定领域开展需求分析，将 TRIZ 分析矩阵知识模型中通用的知识要素（技术问题、技术方案、技术功能、技术效果）延伸细化为特定领域的知识实体和关系类别；在此基础上建立知识实体及其关系与图谱中知识节点和知识关系的映射，制定统一的语义关系分类规范，完成领域知识图谱模式层设计，用于指导后续的领域知识图谱构建和 TRIZ 分析矩阵构建。

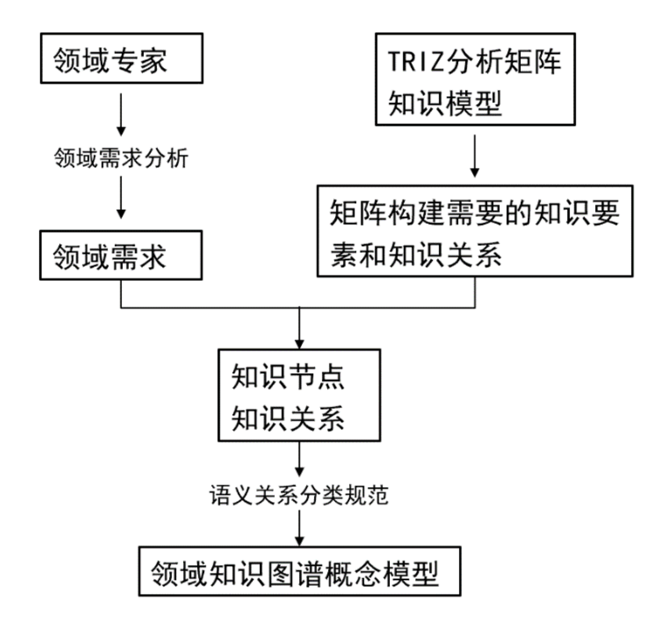


图 2 领域知识图谱模式层设计的技术路线

面向不同领域开展知识图谱模式层设计时，由于领域需求的多样性和差异性，不同领域下技术问题、技术方案、技术功能、技术效果所对应的创新要素也有所区别，由此形成特定领域的 TRIZ 分析矩阵知识模型。后续本文将以薄膜磁头技术领域为例开展面向 TRIZ 分析矩阵的领域知识图谱模式层设计，详见“4.2.1 薄膜磁头技术领域知识图谱模式层设计”。

3.2 面向专利的知识实体和知识关系抽取算法设计

3.2.1 知识实体和知识关系抽取的技术路径

基于 BERT 预训练语言模型技术，研究设计从专利文献中自动识别 TRIZ 分析矩阵构建所需要的知识实体及关系的智能化工具。知识实体和知识关系抽取的技术路径如图 3 所示，包括预训练（Pre-train）、继续预训练（Continual pre-train）、微调（Fine-tuning）3 个阶段。预训练阶段直接引入谷歌利用 12 层的 Transformer Encoder 在 Wikipedia 和 Book Corpus 语料上训练得到的原始 BERT 模型；继续预训练阶段针对特定领域专利文献的实体关系特征，利用领域未标注的大规模专利语料对原始 BERT 模型进行继续预训练，使模型学到更多的专利句法结构知识和特定领域知识，得到特定领域的 BERT 词嵌入；微调阶段利用少量特定任务的标注语料分别对实体识别模型和关系抽取模型进行训练和调优。本研究希望能通过上述技术路径，一方面提升实体识别及关系抽取模型的性能，另一方面减少微调阶段模型对标注数据的依赖。

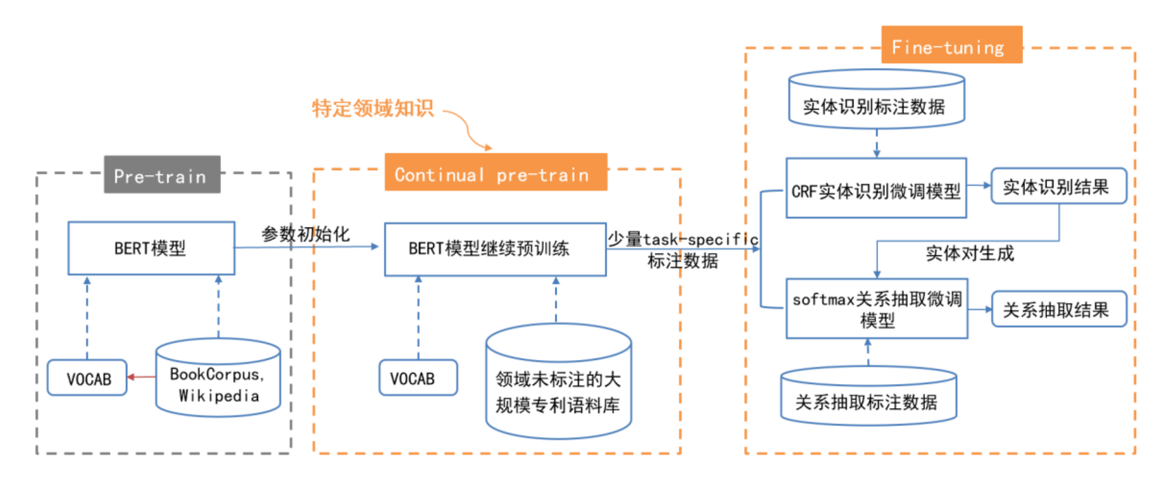


图 3　知识实体和知识关系抽取的技术路径

3.2.2　BERT 模型继续预训练

BERT 模型主要有“预训练+微调”、“预训练+继续预训练+微调”、“重新预训练+微调”三种训练和使用模式（见图 4），ACL2020 Best Paper 提名奖论文《Don’t Stop Pretraining:Adapt Language Models to Domains and Tasks》[24]做了很多语言模型预训练的实验，认为在目标领域的数据集上进行继续预训练可以提升预训练语言模型在处理该领域任务时的效果。陈亮等[25]通过对专利数据集的对比分析发现，不同技术领域的专利数据集之间存在较大差异，并将专利全领域特征的词嵌入和特定领域特征的词嵌入进行了对比，发现特定领域的词嵌入在领域任务中的表现效果更好。

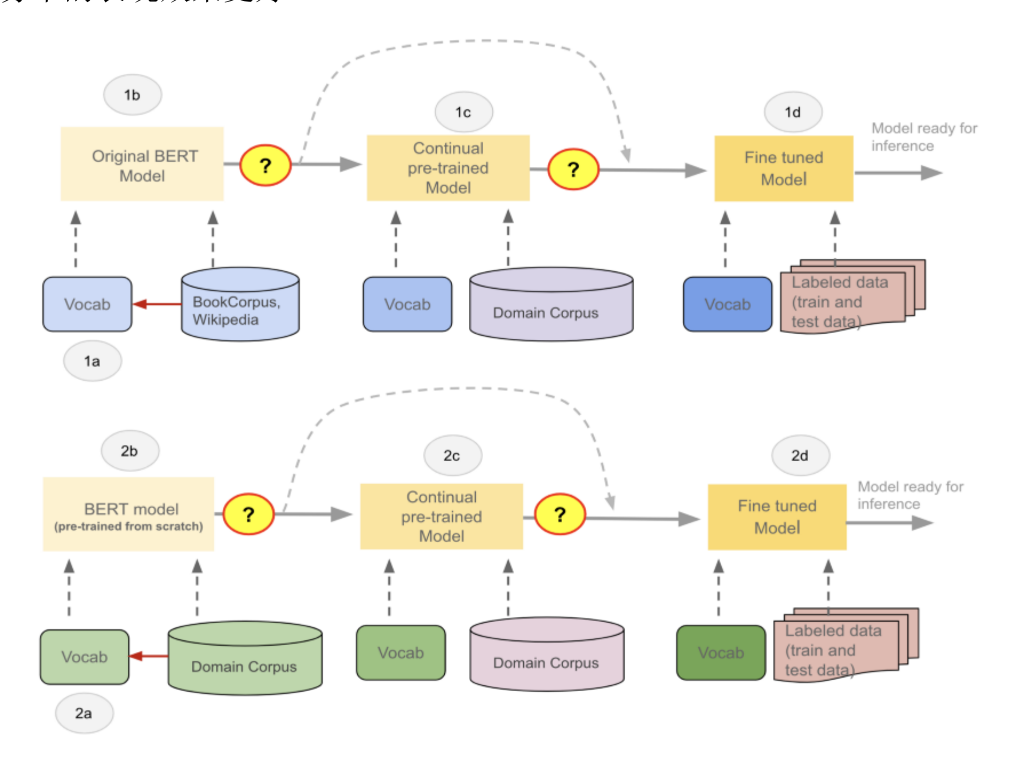


图 4　BERT 模型的训练路径

因此，本研究拟采用 “预训练+继续预训练+微调”的 BERT 模型训练路径。

针对特定领域专利文献的实体关系特征，利用目标领域未标注的专利语料对原始 BERT 模型进行继续预训练，使模型学到更多的专利句法结构知识和特定领域知识，从而得到特定领域的 BERT 词嵌入，以实现更优的实体识别和关系抽取性能。

3.2.3 基于 BERT-CRF 的实体识别模型设计

在序列标注思想的指导下，本文采用 BERT-CRF 实体识别模型，模型架构如图 5 所示。整个模型架构由特征表示、特征编码、标签解码三个部分组成。特征表示步骤由 BERT 模型对输入的文字进行分布式向量表示。特征编码步骤主要对 BERT 提供的输入向量进行变换，通过线性层提取句子的语义特征，将隐状态序列向量维度转换为标注标签数量维度，得到每一个标注标签的预测分值。标签解码步骤使用 CRF 进行解码，将 BERT 层提取到的特征作为输入，然后由 CRF 层负责考虑上下文标签的影响，进而得到使得条件概率最大的实体标签序列。

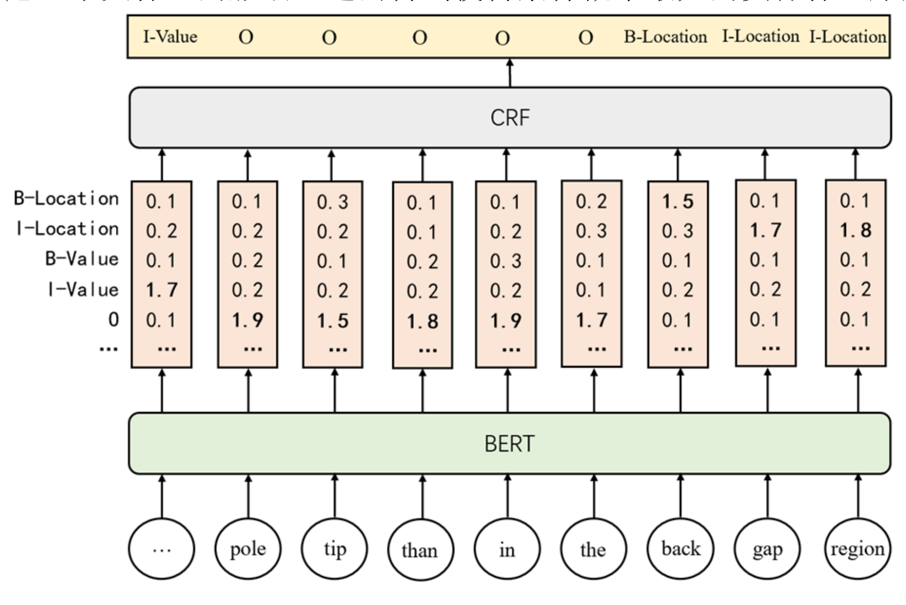


图 5 BERT-CRF 实体识别模型架构

3.2.4 基于 BERT-softmax 的关系抽取模型设计

关系抽取采用多标签分类的思想，构建 BERT-softmax 关系抽取模型。具体的，BERT 负责提供句子的词嵌入表示，学习句子的语义特征，而后将学习到的语义信息接入 softmax 分类器，输出关系分类结果。利用标注语料对关系分类模型进行训练，输入包含实体对的句子，输出关系类别。

此外，由于本研究采用基于流水线的实体识别和关系抽取方法，在实体对生成阶段会迭代生成大量不存在关系的实体对，这些实体对会对关系抽取模型的训练产生干扰。为使关系抽取模型能够更好地学习无关系实体对的特征，在训练过程中，本文为不存在关系的实体对分配“no_relation”的特殊类型，作为负样本添加到训练集当中，在训练过程中与其它关系类型同等看待，以改善关系抽取模型的性能。

3.3 基于领域知识图谱的 TRIZ 分析矩阵自动构建

基于领域知识图谱的 TRIZ 分析矩阵自动构建过程如图 6 所示，包括领域知识图谱构建和 TRIZ 分析矩阵构建两个阶段。

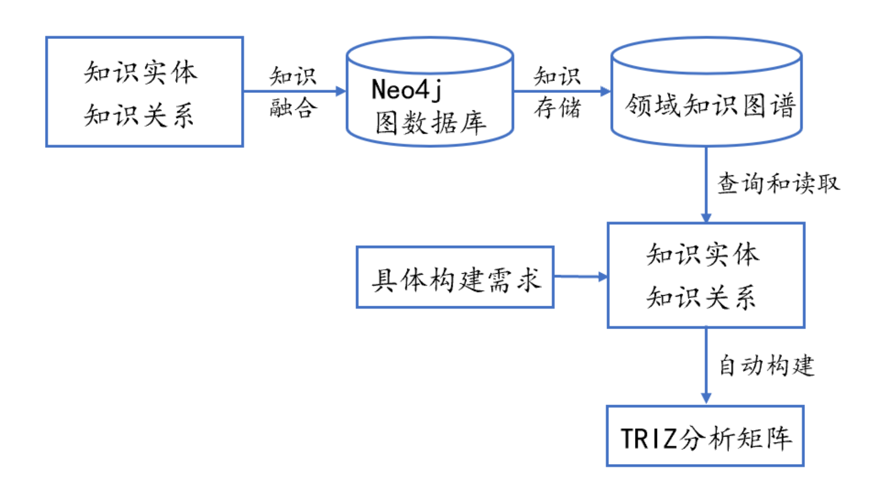


图 6　基于领域知识图谱的 TRIZ 分析矩阵自动构建

首先，对知识实体和知识关系的抽取结果进行简单融合，通过构建实体 ID_Name 对应表保证每一个实体都有唯一的 ID 与之对应，利用实体 ID 对相同的知识实体进行合并，对相同三元组进行剔除。在此基础上，通过图数据库实现实体数据和关系数据的存储，将其转换为数据库中的节点和关系，完成领域知识图谱构建。在利用 Neo4j 图数据库完成领域知识图谱构建的基础上面向具体需求，利用该图数据库进行交互式查询和关联化推理，自动查询和读取所需要的知识实体和关系，实现 TRIZ 分析矩阵的自动构建，为后续的系统问题分析、技术功效分析、技术演化分析、技术路径分析等提供有效支持，加速技术创新。

4　研究实证

计算机硬盘领域的薄膜磁头能显著减少磁头和磁片的距离，增加数据密度，提高准确率，对我国高性能磁记录产业的发展具有重要意义。因此，本文在上述领域通用的 TRIZ 分析矩阵智能构建方法体系下，以薄膜磁头技术领域为例开展实证研究，具体工作包括：1）完成薄膜磁头技术领域专利知识抽取的 BERT 模型构建；2）开展薄膜磁头技术领域 TRIZ 分析矩阵构建实证研究。

4.1　薄膜磁头技术领域专利知识抽取的 BERT 模型构建

4.1.1　实验数据

本实验选用陈亮等[25]构建的 TFH-2020 数据集开展薄膜磁头技术领域专利实体识别和关系抽取实验研究。TFH-2020 是面向硬盘中薄膜磁头技术的带标注的专利数据集。该数据集语料检索自美国专利商标局（USPTO），由薄膜磁头技术领域 1010 篇专利摘要构成，共包含 22742 个实体和 17421 个语义关系。数据集设置了 17 种实体类型规范和 15 种语义关系类型规范，实体类型包括：物质流、信息流、能量流、系统、组件、属性、形状、材料、状态、位置、测量对象、值、科学概念、功能、效果、结果、其它；关系类型包括：实例、别名、位置关系、部分关系、因果关系、构造、属性、以…方式、形成、比较、测量、操作、产生、目的。

4.1.2　BERT 模型继续预训练实验

“磁头”是“薄膜磁头”的上位概念。考虑到专利文献的数量，本文面向计算机硬盘中的磁头领域开展 BERT 模型继续预训练实验。具体的，获取大量磁头领域未标注的专利语料，在 BERT 模型原有参数的基础上进行继续预训练，使模型学到更多的专利句法结构知识和磁头领域知识，训练得到磁头领域的 BERT 词嵌入——BERT-MH。

（1） 用于继续预训练的专利语料

本实验选择 Derwent Innovations Index（DII 德温特创新索引）作为 BERT 模型继续预训练的专利语料来，以“magnetic head”作为检索词，在主题字段中进行检索，共得到 44,705 条专利记录（包含标题和摘要信息），经正则匹配、筛选、清洗后得到每条专利记录的标题和摘要文本，构成磁头领域未标注的专利语料库。最终得到的语料库大小为 48.7MB，共包含 431,210 条句子。

（2）面向磁头领域的 BERT 模型继续预训练

本文采用 PyTorch 库的 Transformers 模型进行 BERT 模型的继续预训练。将上文构建的磁头领域未标注专利语料按 9:1 随机划分为训练集和验证集，采用论文[11]中的建议技巧，从现有的 BERT 检查点开始，在语料上运行其他预训练步骤。继续预训练实验中的部分重要参数设置见表 1。

表 1 面向磁头领域的 BERT 模型继续预训练实验参数设定

参数	参数设定
max_seq_length	128
train_batch_size	32
eval_batch_size	8
learning_rate	5e-5
num_train_steps	100000
num_warmup_steps	10000

整个训练过程在 Linux 服务器 Tesla V100 GPU 上完成。训练完成后，便获得了以.bin 结尾的 PyTorch 版本的磁头领域 BERT 模型。本文将该在磁头领域未标注专利语料上继续预训练得到的 BERT 词嵌入称为 BERT-MH，用于后续的知识实体识别实验和知识关系抽取实验。

4.1.3 知识实体识别实验

在知识实体识别实验中，将 TFH-2020 数据集在文档级别按 6:2:2 的比例随机划分为训练集、验证集和测试集，训练集、验证集和测试集中专利文档和句子的数量分布情况如表 2 所示。

表 2 数据集中专利文档和句子的数量分布

数据集	文档数量	句子数量
训练集	606	2567
验证集	202	878
测试集	202	786

知识实体识别实验共设计了两个实验任务：模型性能对比实验和标注数据依赖性实验，分别用来验证该技术路径在提升模型性能和减少对标注数据依赖性上的有效性。

（1）模型性能对比实验

采用 BERT-softmax、BERT-CRF 典型架构构建实体识别模型，开展对比实验，以实现最优性能。实验要素设置如表 3 所示。为实验得到表现最佳的实体识别模型，将上述不同的实验要素分别进行组合，一共得到 4 个深度学习模型进行实验，模型在实体识别任务中的性能表现如表 4 所示。其中 BiLSTM-CRF+MH-46K 模型是 TFH-2020 专利数据集论文中采用的实体识别模型。

表 3　不同的对比实验要素

	要素 1	要素 2
BERT 预训练语言模型参数	BERT	BERT-MH
下游神经网络	softmax	CRF

表 4　实体识别模型的实体识别效果

模型设置	Precision	Recall	F1
BiLSTM-CRF+MH-46K	78.0%	78.0%	78.0%
BERT+softmax	82.32%	84.00%	83.16%
BERT-MH+softmax	82.76%	83.93%	83.34%
BERT+CRF	84.44%	84.04%	84.24%
BERT-MH+CRF	84.99%	84.87%	84.93%

从实验结果可以看出，BERT-MH+CRF 模型在面向薄膜磁头技术领域的专利实体识别任务中实现了最优性能，证明本研究提出的“预训练+继续预训练+微调”技术路径可以有效提升专利实体识别模型的性能。进一步分析上述实验结果，可以得到以下 3 点结论：1)本实验采用的模型架构与数据集论文中采用的 BiLSTM-CRF+MH-46K 架构相比有较大的性能提升，在一定程度上说明动态词嵌入可以比静态词嵌入学习到更多的特征知识；2）相同下游神经网络模型下，BERT-MH 模型的性能优于 BERT 模型，说明在面向领域任务时，利用特定领域的未标注语料对 BERT 模型进行继续预训练，是提升下游模型性能的一种有效方式；3）相同 BERT 预训练语言模型下， CRF 模型的性能优于 softmax，说明面向具体任务时，依旧可以通过融合其他网络的方式改善模型的性能。

（2）标注数据依赖性实验

基于上述性能表现最优的 BERT-MH+CRF 模型，分别用 50%、40%、30%、20%、10%的标注数据重复上述实体识别实验。在训练过程中，除 epochs 参数有差异外，模型其它参数设置均相同。采用微平均和宏平均两种方式对实体识别结果进行评估，实验结果如表 5 和图 7 所示。

表 5　不同标注数据体量下模型的实体识别效果

标注数据体量	epochs	Micro-average			Macro-average		
		Precision	Recall	F1	Precision	Recall	F1

100%	10	84.99%	84.87%	84.93%	69.27%	68.38%	68.42%
50%	20	83.03%	84.35%	83.69%	58.05%	55.36%	55.50%
40%	50	79.35%	79.66%	79.50%	64.74%	60.19%	60.94%
30%	40	78.63%	79.54%	79.08%	55.62%	52.27%	52.45%
20%	70	79.59%	81.27%	80.42%	56.73%	52.32%	52.77%
10%	250	81.04%	81.22%	81.13%	46.41%	43.81%	44.40%

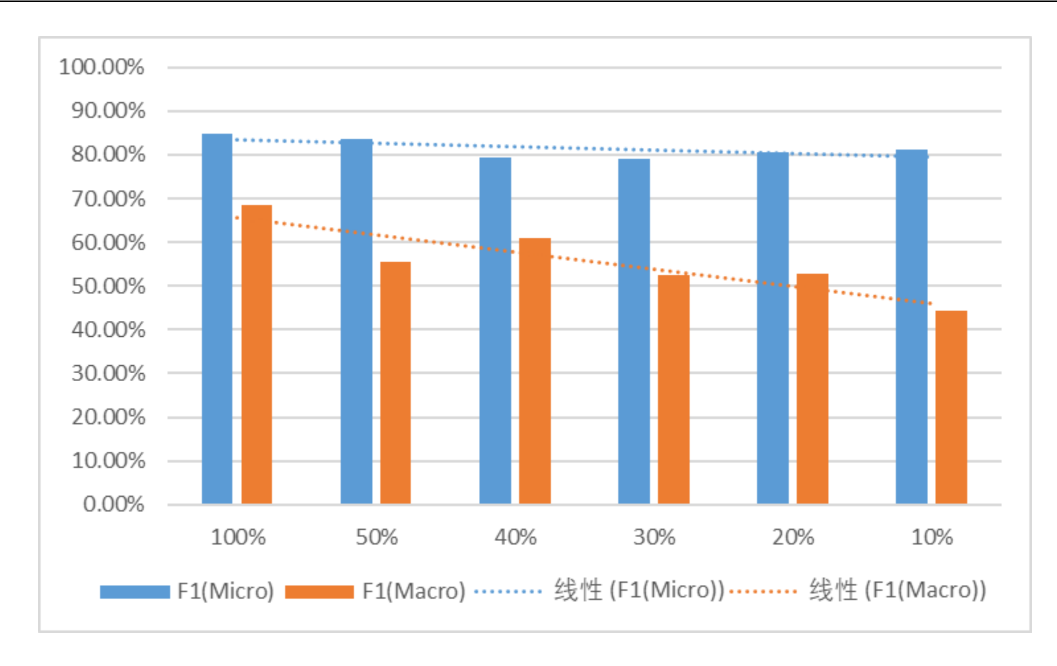


图 7　不同标注数据体量下模型的实体识别效果

从表 5 和图 7 中可以观察到，随着标注数据体量的减小，微平均评估方式下的 F1 值有所下降，但降幅不大；宏平均评估方式下的 F1 值则降幅较大，从 68.42%（全部标注数据）下降到 44.40%（10%标注数据）。这在一定程度上说明当数据集存在严重不均衡的问题时，减少标注数据的体量，对样本量少的实体类别影响较大。但总体上可以认为，在继续预训练阶段利用大规模未标注的领域专利语料对 BERT 模型进行继续预训练，可以在微调阶段减少模型对标注数据的依赖。

4.1.4　知识关系抽取实验

（1）实体对生成

以句子为单位遍历生成实体对（避免实体与自身组合）加入实体对候选集中，再利用实体对生成规则对实体对候选集进行过滤，删除掉显然不可能形成语义关系的实体对；对于筛选后保留的实体对，与句子中的标准关系信息进行匹配，生成正确的关系类型，对于没有任何关系类型标注的实体对，则为其分配没有关系的特殊类型。上述操作后，共生成了 205,119 个用于关系抽取的实体对，具体分布情况如表 6 所示。可以看出，no_relation 类型的实体对共有 183,140 个，占总数的 89.3%，关系类型的样本分布极端不平衡，这给关系抽取模型的训练带来了极大的挑战。

表 6　数据集中实体对的数量分布

数据集	实体对总数	no_relation 类型的实体对数量
训练集	122,890	110,873
验证集	40,423	36,161
测试集	41,806	36,106

实验采用 BERT-softmax 典型架构构建关系抽取模型。BERT 预训练语言模型选择原始 BERT 模型和继续预训练得到的磁头领域 BERT-MH 模型。将 BERT 模型和 BERT-MH 模型分别与 softmax 分类器进行组合，以期实验得出最佳的关系抽取模型。经过训练后，模型在测试数据集上的性能表现如表 7 所示。其中 BiGRU-HAN 模型是 TFH-2020 专利数据集论文中所采用的关系识别模型。

表 7　关系抽取模型的整体评估结果

	Precision	Recall	F1
BiGRU-HAN *with no_relation*	87.9%	87.9%	87.9%
BiGRU-HAN *without no_relation*	41.5%	41.5%	41.5%
BERT+softmax *with no_relation*	89.15%	89.15%	89.15%
BERT+softmax *without no_relation*	63.56%	63.56%	63.56%
BERT-MH+softmax *with no_relation*	89.70%	89.70%	89.70%
BERT-MH+softmax *without no_relation*	63.70%	63.70%	63.70%

表 7 中的“with no_relation”行是同时考虑“no_relation”类别而获得的评测数据，在 BERT-MH+softmax 模型上 F1 值达到了 89.7%。但是，本文真正关注的是表 7 中“without no_relation”行的评测结果，它们更真实地反映了模型在专利关系抽取任务中的性能。可以看到，BERT-MH+softmax 模型在该任务中取得了最优的性能，F1 值为 63.7%，略高于 BERT+softmax 模型的 63.56%，大幅高于 BiGRU-HAN 模型的 41.5%。这说明：采用“预训练+继续预训练+微调”的技术路径可以有效提升关系抽取模型的性能；与通用领域的 BERT 词嵌入相比，特定领域的 BERT 词嵌入在面向领域的任务中能更好地提升下游模型的性能。

同时值得一提的是，BERT-MH+softmax 关系抽取模型的 F1 表现仍没有达到关系抽取模型的上游水平，可能的原因如下：1）专利文本中包含的实体要比通用文本多得多，导致无关系实体对的比例比普通文本大得多，给模型训练带来难度。2）流水线方法的错误传播问题，错误的实体识别结果不可避免地会导致语义关系抽取的错误，降低模型性能。

4.2　薄膜磁头技术领域 TRIZ 分析矩阵构建实证研究

应用上述薄膜磁头技术领域专利知识抽取模型——BERT-MH+CRF 实体识别模型和 BERT-MH+softmax 关系抽取模型，开展薄膜磁头技术领域 TRIZ 分析矩阵构建实证研究。

4.2.1　薄膜磁头技术领域知识图谱模式层设计

薄膜磁头技术专利的主要内容通常是关于磁头的系统结构、工作机制及其组成部分的位置、属性、材料构成等。因此，对应于薄膜磁头技术领域，技术问题通常指专利的研究对象，包括物质流、信息流、能量流；技术方案通常指薄膜磁头的系统结构，包括系统中组件的属性、形状、材料、位置等；技术功能指磁头系统所实现的功能；技术效果指系统的影响、效果，以及所产生的结果。

在此基础上，本文参考 TFH-2020 数据集中定义的实体和关系类型，明确薄膜磁头技术领域知识图谱的实体类别为：物质流、信息流、能量流、系统、组件、属性、形状、材料、位置、功能、效果、结果，共 12 类知识实体。知识实体与技术问题、技术方案、技术功能、技术效果 4 大知识要素的对应关系见表 8。明确薄膜磁头技术领域知识图谱的关系类别为：别名、位置关系、部分关系、因果关系、构造、属性、以……方式、形成、操作、目的，共 10 类知识关系。最终构建得到的薄膜磁头技术领域知识图谱概念模型如图 8 所示。

表 8　语义功能类型与实体类型的对应关系

语义功能类型	实体类型
技术问题	物质流、信息流、能量流
技术方案	系统、组件、属性、形状、材料、位置
技术功能	功能
技术效果	效果、结果

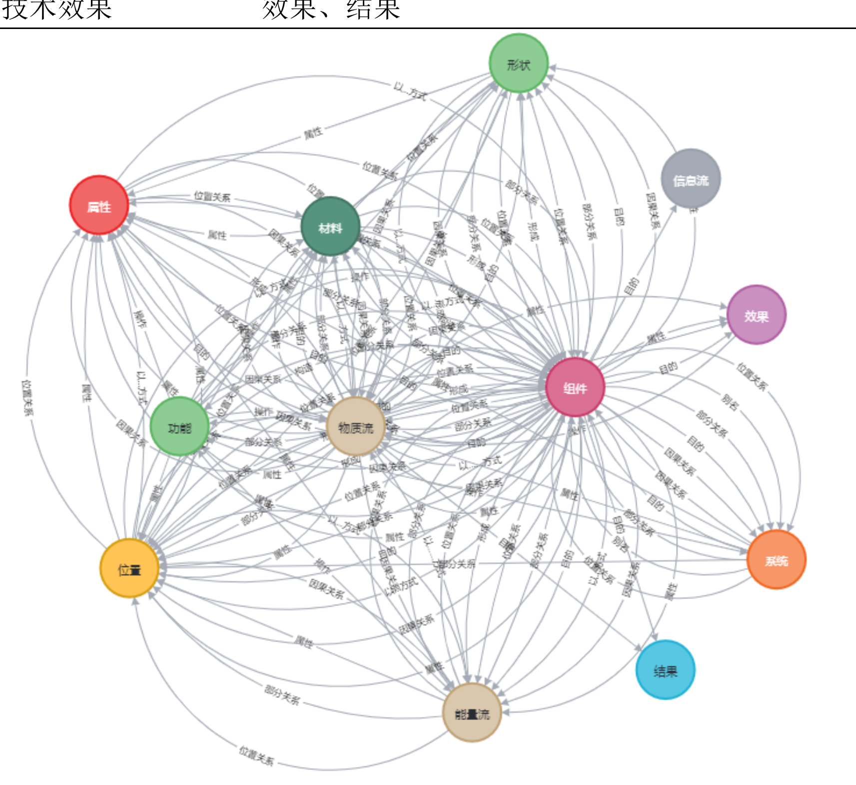


图 8　面向 TRIZ 分析矩阵的薄膜磁头技术领域知识图谱概念模型图

4.2.2　薄膜磁头技术领域专利数据获取

在德温特创新索引数据库中以“thin film magnetic head”（薄膜磁头）作为检索词，在标题字段中进行检索，共检索得到 1732 篇与薄膜磁头技术相关的专利

记录（包含标题和摘要信息）。经数据清洗和统计后，得到文件大小为 2.88MB 的专利数据文档，作为薄膜磁头技术领域 TRIZ 分析矩阵构建的专利语料。该语料共包括 15,666 条有效句，372,650 个词，句子平均长度为 23.8 个词。

4.2.3 知识实体识别和关系抽取结果

完成数据预处理后，调用 BERT-MH+CRF 模型对薄膜磁头技术领域专利语料开展实体识别，识别共得到 70,844 个实体，实体类别的数量分布如表 9 所示。完成实体识别后，共生成 335,596 个实体对。关系抽取实验调用 BERT-MH+softmax 模型进行，识别得到的关系类别数量分布如表 10 所示。

表 9　实体类别的数量分布

实体类型	数量	实体类型	数量
物质流	422	位置	7881
信息流	352	形状	1587
能量流	3456	材料	6434
系统	2279	功能	3591
组件	37008	效果	1585
属性	5995	结果	254

表 10　关系类别的数量分布

关系类型	数量	关系类型	数量
位置关系	9243	因果关系	914
属性	7220	以……方式	792
目的	2379	构造	1598
部分关系	9401	形成	427
操作	1623	别名	597
no_relation	301402		

4.2.4 薄膜磁头技术领域知识图谱构建

对上述知识实体和知识关系的抽取结果进行简单融合，通过构建实体 ID_Name 对应表保证每一个实体都有唯一的 ID 与之对应，简单去重后得到的实体数量为 14,014 个，进一步剔除“no_relation”关系后，得到的有效关系数量为 20,863 条。

将去重后得到的实体集合和关系集合结构化保存并赋予相应标签，录为 CSV 格式文件存储。利用 Neo4j 图数据库提供的 neo4j-admin import 工具批量导入 CSV 格式的实体文件和关系文件，成功导入 14,014 个实体节点和 20,863 个关系，构建得到的薄膜磁头技术领域知识图谱如图 9（左）所示。可以发现，薄膜磁头技术领域知识图谱整体呈现聚集状，只有少数实体关系分散在周围。图 9（右）则展示了薄膜磁头技术领域知识图谱的实体和关系结构细节。

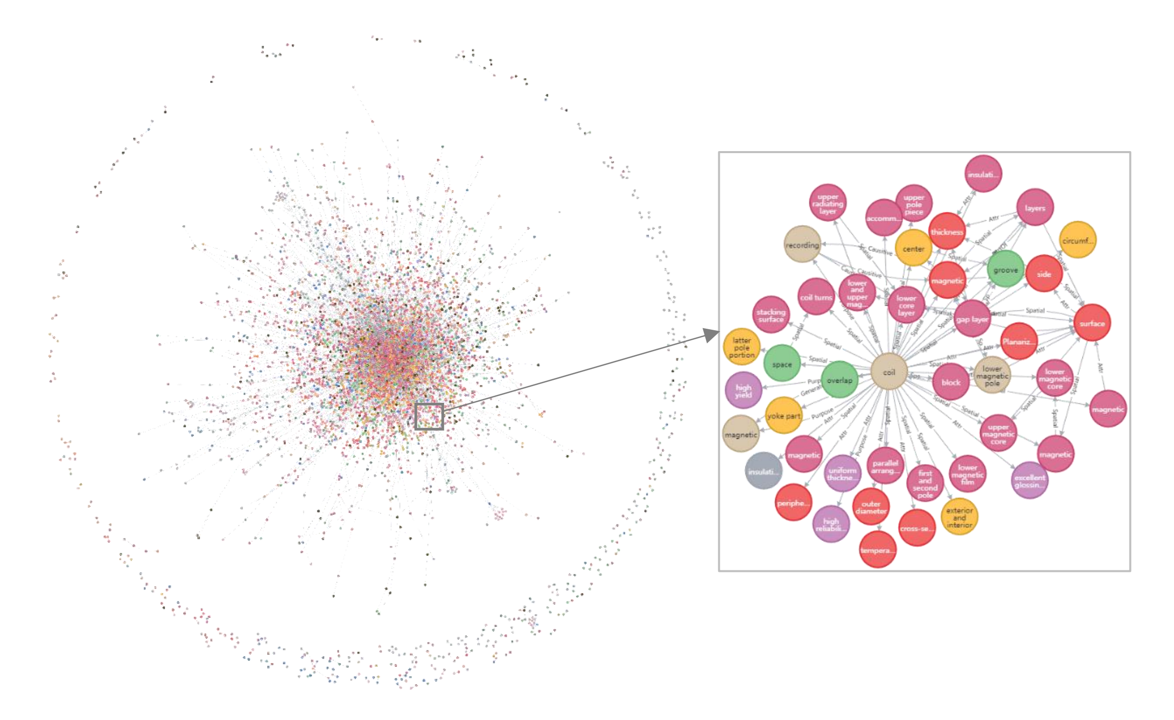


图 9　薄膜磁头技术领域知识图谱

4.2.5　薄膜磁头技术领域 TRIZ 分析矩阵构建

记录密度和频率是评价薄膜磁头性能的重要指标，基于此，本文围绕“如何提高薄膜磁头的记录密度和频率”这一技术创新需求开展 TRIZ 分析矩阵构建示范。首先，明确该技术创新需求追求的技术功效为薄膜磁头的高记录密度和频率。因此，以“high recording density and frequency”（高记录密度和频率）作为关键词，在构建得到的薄膜磁头技术领域知识图谱中进行查询，查询语句为：“match(x) where x.name =' high recording density and frequency' return x”。

查询得到“高记录密度和频率”实体，发现“发明的薄膜磁头”实体与该实体直接关联，进一步扩展与“发明的薄膜磁头”实体相关联的实体，发现“宽度”、“磁极部分”、“轨道变窄”、“示意结构”实体与其直接相关，由此推断磁极部分是该薄膜磁头实现高记录密度和频率的关键组件，于是继续拓展与“磁极部分”实体相关联的实体。拓展后发现关联实体较多，于是人工对关联实体的相关程度进行判别，删除关联性较小的实体，保留重要实体。经上述关系扩展和数据规范操作后，最终得到的查询结果如图 10 所示。

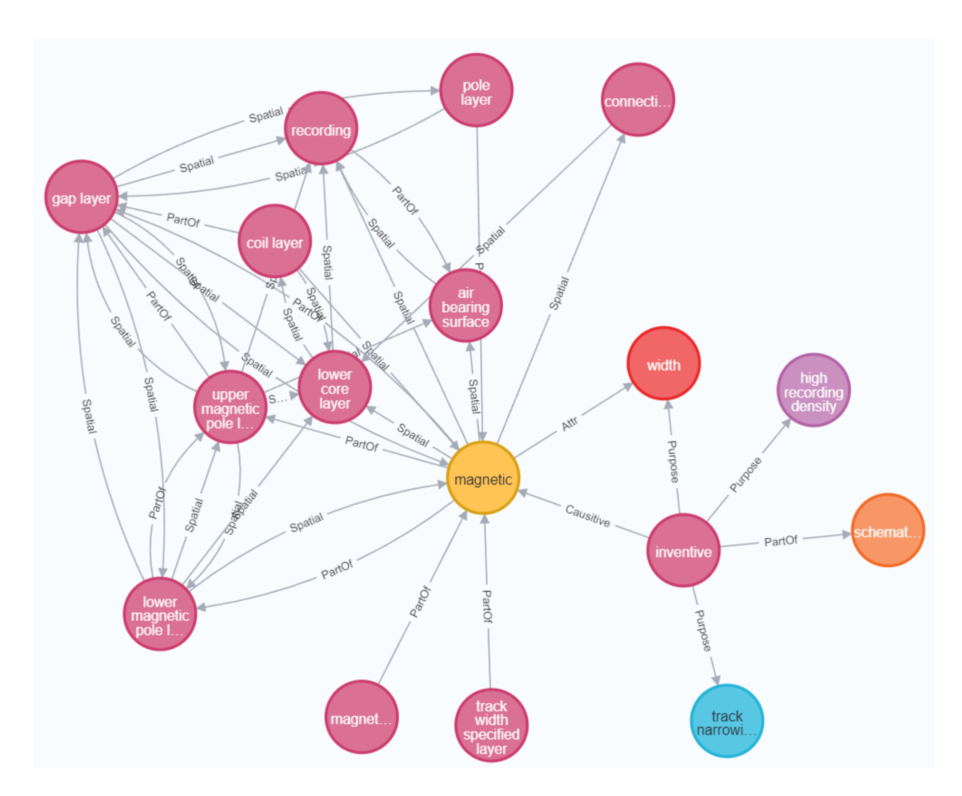


图 10　薄膜磁头技术领域知识图谱查询结果

从查询结果中可以看出，“高记录密度和频率”实体与“发明的薄膜磁头”实体直接相关，“发明的薄膜磁头”实体又与“示意结构”、“磁极部分”、“宽度”、“轨道变窄”等实体直接相关。因此，本文面向上述重要知识实体和关系构建 TRIZ 分析矩阵，构建结果如表 11 所示。

表 11　薄膜磁头技术领域 TRIZ 分析矩阵

	发明的薄膜磁头（组件）	示意结构（系统）	磁极部分（组件）	宽度（属性）	轨道变窄（结果）	高记录密度和频率（效果）
发明的薄膜磁头（组件）		部分关系	部分关系	目的	目的	目的
示意结构（系统）	部分关系					
磁极部分（组件）	部分关系			属性		
宽度（属性）	目的		属性			
轨道变窄（结果）	目的					
高记录密度和频率（效果）	目的					

对上述 TRIZ 分析矩阵进行分析，可以发现“高记录密度与频率”的效果和

“轨道变窄”的结果、“宽度”属性同时出现，而上述目的的实验实现又与“薄膜磁头”的“磁极部分”组件相关，这提示高记录密度与频率可能与磁极部分轨道的宽度有关，减小薄膜磁头磁极部分轨道宽度可以作为提高磁头记录密度和频率的探索方向之一。

5 结语

本文针对传统人工构建系统相互作用分析矩阵存在的速度慢、代价高、知识要素不足等问题，提出 TRIZ 分析矩阵智能构建方法研究。在 TRIZ 创新理论的指导下，面向特定领域拓展系统相互作用分析矩阵的知识内容，定义知识实体和关系类别，完成面向 TRIZ 分析矩阵的领域知识图谱模式层设计；基于“预训练+继续预训练+微调”的技术路径，采用 BERT 预训练语言模型技术，从专利文献中自动识别知识实体和关系，构建领域知识图谱；面向具体需求，从领域知识图谱中自动查询和读取所需要的知识实体和关系，最终实现 TRIZ 分析矩阵的自动构建。经实证验证，所提出的方法可以实现 TRIZ 分析矩阵自动构建的目标，可以为技术创新提供有效的知识支持。

但本研究仍存在一定的局限性：（1）本文基于流水线方法开展实体识别和关系抽取实验，存在错误传播问题等问题；（2）本文在基于领域知识图谱构建 TRIZ 分析矩阵的过程中，主要是基于交互式查询的方式获取所需的知识实体和关系信息，缺乏知识推理技术的应用，无法揭示实体间潜在的知识关联；（3）本文构建的 TRIZ 分析矩阵在质量上仍存在不足。未来将引入知识推理技术，深入挖掘图谱中实体间潜在的知识关联，提高 TRIZ 分析矩阵的质量，为技术创新提供更多线索。

参考文献：

[1] 曾蓓，吕建秋. 基于 CiteSpace 的国内 TRIZ 研究热点及前沿分析[J]. 科技管理研究, 2019, 39(18): 260-265.

[2]Kiela D, Guo Y, Stenius U, et al. Unsupervised discovery of information structure in biomedical documents[J], 2015, 31(7): 1084-1092.

[3] Guo Y, Silins I, Stenius U, et al. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review[J], 2013, 29(11): 1440-1447.

[4] Guo Y, Reichart R, Korhonen A. Improved information structure analysis of scientific documents through discourse and lexical constraints[C]. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013: 928-937.

[5] Mccallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[J], 2003.

[6] Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition[C]. COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[7] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J], 2016.

[8] Yoon W, So C H, Lee J, et al. Collabonet: collaboration of deep neural networks for biomedical named entity recognition[J], 2019, 20(10): 55-65.

[9] 余丽，钱力，付常雷，等. 基于深度学习的文本中细粒度知识元抽取方法研究[J]. 数据分析与知识发现, 2019, 3(01): 38-45.

[10] Gao H, Gui L, Luo W. Scientific literature based big data analysis for technology insight[C]. Journal of

Physics: Conference Series, 2019: 032007.
[11] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J], 2018.
[12] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J], 2020, 36(4): 1234-1240.
[13]Zhang X, Zhang Y, Zhang Q, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods[J], 2019, 132: 103985.
[14]Wu S, He Y. Enriching pre-trained language model with entity information for relation classification[C]. Proceedings of the 28th ACM international conference on information and knowledge management, 2019: 2361-2364.
[15]唐晓波，刘志源．金融领域文本序列标注与实体关系联合抽取研究[J]．情报科学, 2021, 39(05): 3-11.
[16]Tsourikov V M, Batchilo L S, Sovpel I V. Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (SAO) structures: Google Patents, 2000.
[17]胡正银，刘春江，隗玲，等．面向 TRIZ 的领域专利技术挖掘系统设计与实践[J]．图书情报工作, 2017, 61(01): 117-124.
[18] Kim H, Hyeok Y, Kim K. Semantic SAO network of patents for reusability of inventive knowledge[C]. 2012 IEEE International Conference on Management of Innovation & Technology (ICMIT), 2012: 510-515.
[19]李晓曼，张学福，宋红燕，等．专利文献技术要素识别方法研究——以纳米肥料领域为例[J]．图书情报工作, 2020, 64(06): 59-68.
[20]李卫超．面向专利的功能信息抽取方法的研究[D]．河北工业大学, 2013.
[21]Nanba H, Kondo T, Takezawa T. Automatic creation of a technical trend map from research papers and patents[C]. Proceedings of the 3rd international workshop on Patent information retrieval, 2010: 11-16.
[22] 赖英旭，李亚娟，刘静．基于本体的水稻育种方法应用知识库构建[J]．北京工业大学学报, 2019, 45(12): 1181-1191.
[23]张娴，胡正银，茹丽洁，等．专利技术供需信息关联知识组织模式研究[J]．图书情报工作, 2016, 60(08): 118-125.
[24] Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks[J], 2020.
[25] Chen L, Xu S, Zhu L, et al. A deep learning based method for extracting semantic information from patent documents[J], 2020, 125(1): 289-312.